Application of Digital Fingerprinting: Duplicate Image Detection

Ashutosh Maharana

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India

Application of Digital Fingerprinting: Duplicate Image Detection

Thesis submitted in

May 2016

to the department of

Computer Science and Engineering

of

National Institute of Technology Rourkela

in partial fulfillment of the requirements

for the degree of

Master of Technology
(Dual Degree)
In

Computer Science and Engineering

by

Ashutosh Maharana
[Roll No. 711CS2034]

under the guidance of

Dr. Ruchira Naskar

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India

Department of Computer Science & Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India. www.nitrkl.ac.in

NIT,Rourkela

May 20, 2016

Certificate

This is to certify that the work in the thesis entitled ***Application of Digital Fingerprinting: Duplicate Image Detection*** by **Ashutosh Maharana**, bearing Roll No. 711CS2034, is a record of an original research work carried out by him under my supervision and guidance in partial fulfilment of the requirements for the award of the Degree of ***Master of Technology (Dual Degree)*** **in** ***Computer Science and Engineering***. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Prof. Ruchira Naskar
Asst. Professor
CSE department of NIT Rourkela

Dedicated to,

My Parents and Teachers.

ACKNOWLEDGEMENTS

First of all, I would like to express my deep sense of respect and gratitude towards my supervisor Prof. Ruchira Naskar, who has been the guiding force behind this work. I want to thank her for introducing me to the field of Multimedia Security and giving me the opportunity to work under her. Her undivided faith in this topic and ability to bring out the best of analytical and practical skills in people has been invaluable in tough periods.

I also want to thank all PhD researchers in NIT Rourkela for making our lab such an incredible work environment.

Besides my thesis advisor I extend my sincere thanks to Prof. Santanu K. Rath, Head of the Computer Science and Engineering Department, and all other faculties of the department for their timely co-operations during the project work. Such a large number of individuals have contributed to my research work, and it is with awesome joy to take the chance to express gratitude toward them. I apologize, in the event that I have overlooked anybody.

Ashutosh Maharana

ABSTRACT

Identifying the content automatically is the most necessary condition to detect and fight piracy. Watermarking the image is the most basic and common technique to fight piracy. But the effectiveness of watermark is limited. Image fingerprinting provides an alternate and efficient solution for managing and identifying the multimedia content. After registering the original image contents, by comparing the colluded image with the original one, the percentage of distortion can be calculated. In this paper presented are one such fingerprinting-based forensic application: Duplicate image detection. To authenticate image content perceptual hash is an efficient solution. Perceptual hashes of almost similar images or near duplicate images are very similar to each other making it easier to compare images unlike cryptographic hashes which vary very radically even in the case of small distortions. Potential applications are unlimited including digital forensics, protection of copyrighted material etc. However, conventional image hash algorithms only offer a limited authentication level for the protection of overall content. In this work, we compared and contrasted different perceptual hashes and proposed a image hashing algorithm which is an excellent trade off of accuracy and speed.

Keywords: Digital Fingerprint, Image Hashing Algorithm, Perceptual Hash

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Fingerprinting of digital media is a technology by which the copyrighted media content can be controlled better by the rightful content owners by effectively identifying, indexing and detecting the colluded media via different media distribution channels and by tracking colluding piracy.

Basically, any hash value that has been extracted from the original media file, has enough characteristics and details that represent the characteristics of original media file and can be identified as a unique representation while comparing to other hash values.[1] Fingerprinting calculation use assortment of content media file properties like casing bits, movement and music changes, camera cuts, splendor level, object developments and so on to spoke to fingerprints which are put away as resource references in database vault. They can distinguish at whatever point a variation of sound/image content originating from various sources is analyzed even in instances of content media file adjustment or modification. The fingerprint calculation deals with balancing between perfect measure of information catch to empower granular examination over the length of content while keeping fingerprints lightweight for reasonable access, indexing, inquiry, and storing of the fingerprints, also commonly known as hashes. Some of the best products in market todays using digital fingerprints to use in digital forensics are iPhaarro, Audible Magic etc.

1.1 IMPORTANTANCE OF DIGITAL FINGERPRINTING

In the modern digital world, new technology and internet have significantly reduced the effort, time and cost for producing, storing and distributing the copyrighted digital content. On the other hand connectivity of internet and new technologies have also enabled digital piracy practitioners to copy and distribute the same content globally. Copyright Infringement can show in numerous ways e.g. tearing substance from the CD/DVD, catching digital content from Television, conveying digital duplicates over long range interpersonal communication and P2P systems, video recording from Theaters, content replicating and copying and so on.[6]

Digital Fingerprinting empowers distributers (and performing rights social orders) to procure advantages in benefits by growing and amplifying the estimation of their digital content through more current economical adaptations like production of digital content on paid video gateways, informal communities destinations creating publicizing incomes, execution and playback of substance crosswise over radio, TV and limited time occasions, utilization of digital content in commercials, tv and radio projects and promoting similar exercises among others.

1.2 DIGITAL FINGERPRINTING VERSUS WATERMARKING

Although both digital fingerprinting and watermarking have almost similar purpose,the characteristics of both technologies are very different from one another. Watermark is a marking or a symbol embedded within the digital content which is perfectly visible. Watermarks serve to control the unlawful spread of digital media by inconvenience more than brilliant innovation.

A watermark is a logo or other recognizing checking set on an image or image that is noticeable at all times. The watermark expects to dishearten Internet clients from taking a photo or an image from one Web webpage and utilizing it for their own motivations without recognizing the source. But there is no real promise that the watermarking technique will work effectively. Digital piracy can still be done by cropping the watermark. Watermarked content can also be redistributed woth or without hiding the source or the rightful content owner.Another form of watermarking(embedding constraints) is used by content owners which helps them to track and catch the colluders. Digital fingerprinting offers a much additionally encouraging approach to limit the spread of copyrighted material. Table no 1 shows the basic differences between fingerprinting and watermarking.

Fingerprints vs. Watermarks	Fingerprint	Watermark
Content remains unchanged	Yes	No
Unique	Yes	Not necessarily
Can be removed	No	Yes
Used retrospectively	Yes	No
Survive processing of content	Yes	Limited
Compatible with all devices	Yes	No
Print stored separately	Yes	No
Propriety	Yes	Yes

Table 2 : Digital Fingerprinting vs. Watermarking

1.3 WORKFLOW FOR DIGITAL FINGERPRINTING

Generating fingerprint and distinguishing and identifying digital content is a vital part of the digital media distributor's media work process with capacity to recognize, track, screen and adapt their digital media. It engages distributers with innovation to anticipate copyright encroachment, gives intends to reach out due monetary advantages to legitimate substance proprietors and block themselves from lawful liabilities emerging because of unlicensed spread of copyrighted material (DMCA rules). A general Digital Fingerprinting process includes content proprietors/studios enrolling their digital media for fingerprinting and making reference digital representation of their contents which is utilized for future examinations. The main steps of digital fingerprinting are:

- Calculating the hash value of digital media content (also known as digital fingerprints) and adding them to the fingerprint database after indexing each and every content.
- Identifying duplicate or colluded media file by checking and comparing the fingerprint with the registered fingerprints in the fingerprint database.
- Taking proper steps to media file depending on the comparision result with the database under copyright laws.

1.4 THESIS ORGANIZATION

The remaining part of the thesis is organised as follows. Chapter 1, as we have seen contains introduction and covers all the basic need to know. Chapter 2 summarizes some of the

related works done on the same field before. Chapter 3 contains proposed duplication image copy detection method using digital fingerprinting. Chapter 4 contains results and discussions on this thesis. And finally in chapter 5 conclusion and some future work is proposed.

Chapter 2

BACKGROUND

Identifying features are extracted from an image by digital fingerprinting (fingerprints) which are unique to each image for a particular hashing algorithm. There are several image hashing or fingerprinting techniques to find the unique value of the image.

2 .1 TYPES OF FINGERPRINTING TECHNIQUES

Identifying features are extracted from an image by digital fingerprinting(fingerprints) which are unique to each image for a particular hashing algorithm. There are several image hashing or fingerprinting techniques to find the unique value of the image. Some of them are semantic (abnormal state examination) and different ones are more flag based (low level analysis). Low level procedures are isolated in two classes: local and global. The later one is quicker however experiences crash and absence of robustness against solid colluding attacks, for example, cropping, impediment, and expansion.[11] The local hashing approach however is more robust to different colluding attacks and image distortion. The extricated components are repeatable, between reliant and more discriminant than the global hashing algorithms. Among local image hasing approaches, the procedures in view of purposes of interest take care of the spatial synchronization issue, as a model of attacking distortion can be assessed from a mapping of such focuses in unique and duplicate images. (As shown in Figure 2.1) Local image fingerprints are fundamentally utilized as a part of content identification and biometry. In a

biometric identification system, the "unique mark" highlights local image key points. Key feature points of a given client can be utilized to integrate a fingerprint. Security and yield bit length are in this manner the fundamental shortcomings of fingerprinting.. Perceptual hash calculations are intended to beat these security issues.

2.2 PROPERTIES OF FINGERPRINTS

The image fingerprints generally need to satisfy the following properties:

a) Robustness: The fingerprints extracted from a degraded image must be identical or nearly identical to fingerprints of the copyrighted image.

b) Pairwise independence: When two different images are considered, the fingerprint extracted from those images should also be different.

c) Database search efficiency: For large-scale database applications, the fingerprints should be efficient for DB search.

2.3 PERCEPTUAL HASHES

A compressed version of data called message digest is calculated in digital fingerprinting by using a hash function. However cryptographic hash functions are totally unacceptable for image hashing as any small distortion in input image will give a very different hash value from the original hash. To encounter these problems, perceptual hash is use in image hashing. Perceptual image hash algorithms use fingerprinting procedures with cryptosystem-like constraints. The properties of a perceptual hash capacity are: ease of calculation, feeble crash resistance, and a hash value of fixed length perceptual image hash. This is more robust to image

distortions. A small change in image will result in a very little change in image hash value. And a large change in image will change the hash value by greater bits.

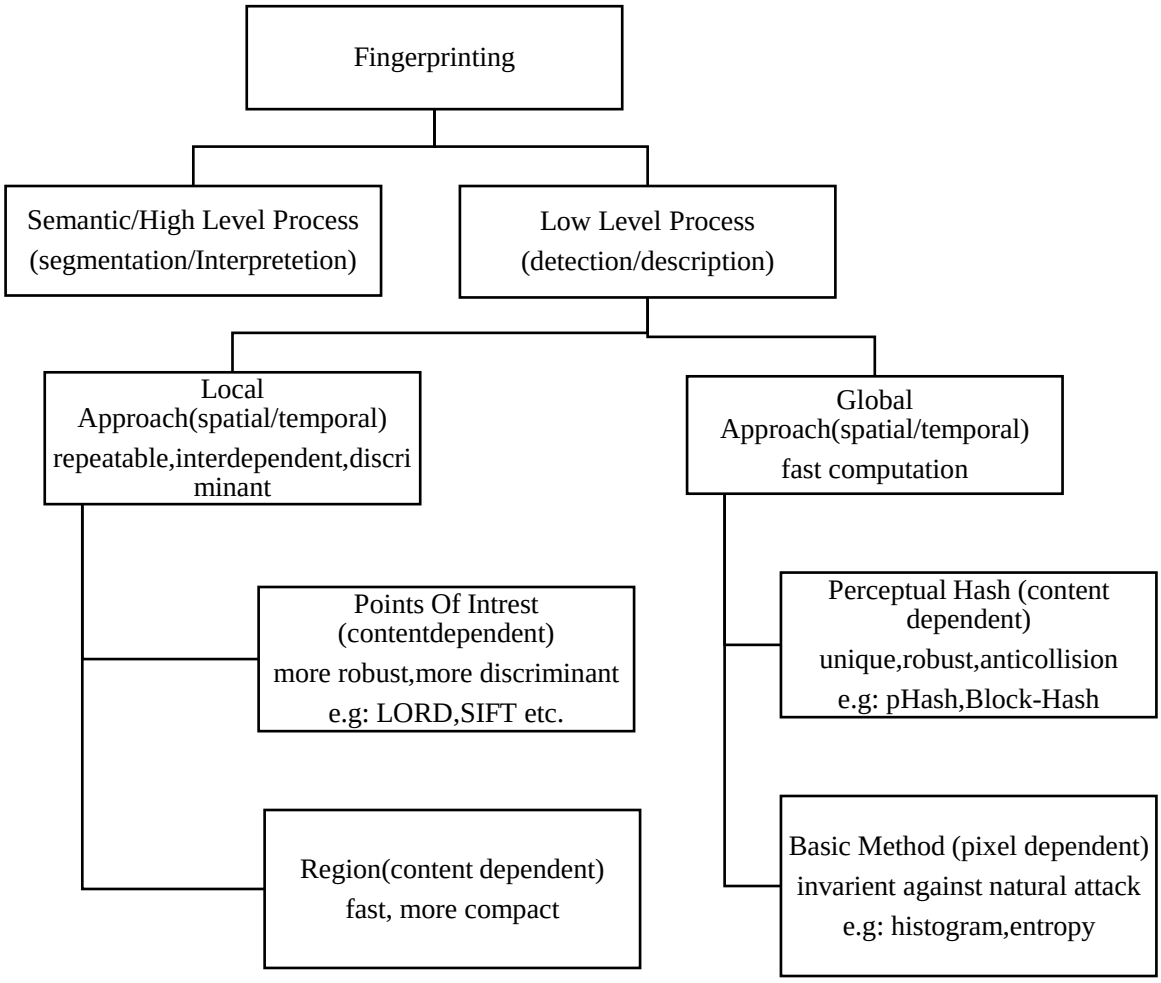

Figure 2.1 : Characterization of different fingerprinting techniques

2.4 HAMMING DISTANCE

Hamming distance is used to compare and contrast two different hashes. The hamming distance denotes the number of bits by which two images are different from one another. A short hamming distance will denote that the two images are almost similar. For an exact copy of the image, the hamming distance will be 0.In Short, if d(x,y) is the hamming distance then d(x,y)is the number of bits by which x and y differ. This can be calculated using XOR ing x and y.

2.5 NORMALIZED HAMMING DISTANCE

The normalized hamming distance can be defined as the average of all the hamming distances to all its corresponding pixels. The smaller the value of the normalized hamming distance better is the robustness of the algorithm. It's only applicable for 2D matrixes or binary hash values.

2.6 RELATED WORK

The framework can just recognize record that it knows; along these lines, another document has first to be enrolled into the framework. (Figure 2.2) So two symmetrical procedures are in this manner required:

1. Indexing the new dataset
2. Detecting the duplicate image

- To enlist another record, we present the first document to the fingerprint extractor, which produces the fingerprint . The fingerprints are indexed in a database together with the name of the record and discretionary metadata.

- To detect the copies of the test image file, we present the test record to the fingerprint extractor to produce its fingerprint. The system recovers them and scans the database for the potential applicant; the reaction is either the name of the document, or a unique message. [4]

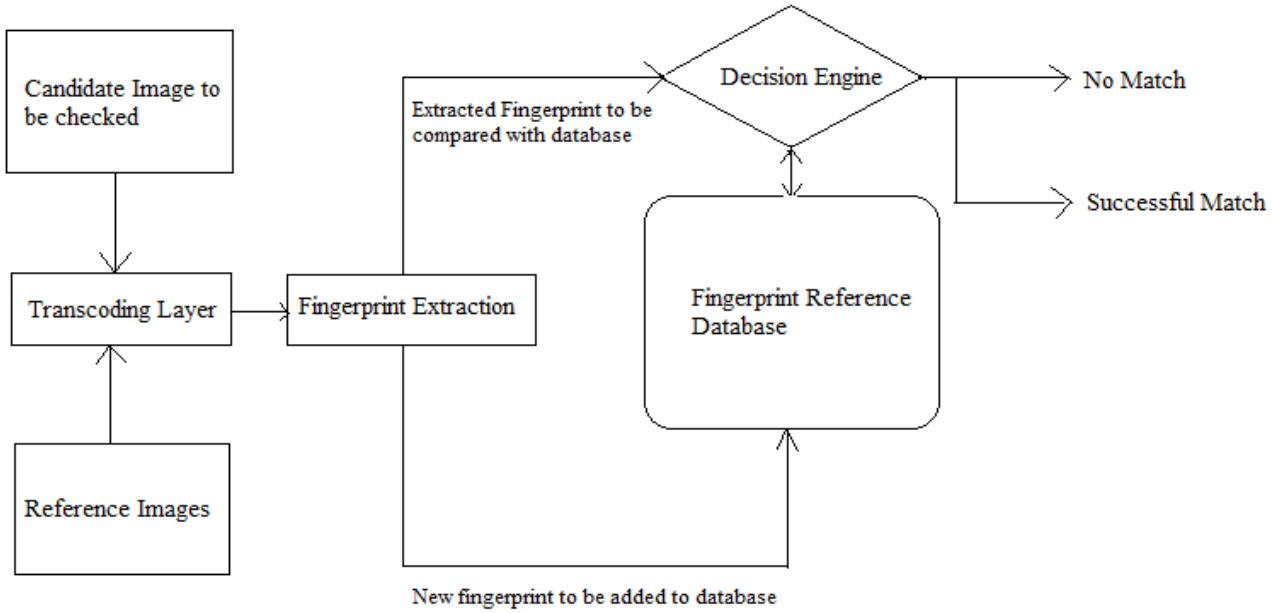

Figure 2.2 : Indexing and Detection in a typical digital fingerprinting system

The system does not require the complete content: only a short section is adequate to recognize the content. Hence, a submitted test may begin anyplace in the opus.The proposed calculation was tuned to give a decent speed versus exactness tradeoff for detecting similar hashes. The method uses a visual hashing algorithm and a feature based local image hash algorithm. At first the test file was analyzed through global approach. The initial step is a

worldwide investigation of the test record. The input image file's hash value is calculated. Then the hash value is compared with the query hashes. Depending on the result of the comparison either the input file is discarded (if duplicate) or indexed in the database as a new image.(Figure 2.3)

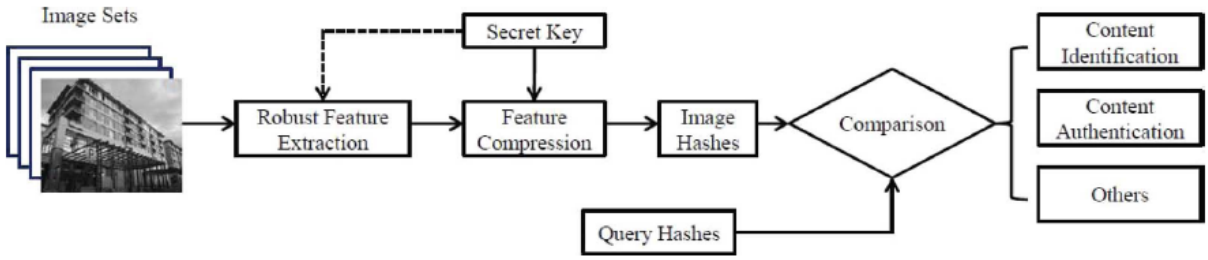

Figure 2.3 Image hash calculation and comparison

The comparability between two contents is normally in light of twofold distance between two components. If there should arise an occurrence of a fingerprinting application, two contents can have two comparable fingerprints with a critical double distance. Paired distance or Hamming distance are in this manner not adjusted to fingerprint applications. In the writing we discover fingerprints of measurement 128, 144 ,180, and so forth. It implies that the fingerprint coordinating application needs to seek a component of measurement 128, 144 or 180 in a database populated by N components of measurement 128, 144 or 180. With the calculation portrayed in, 315 hours of expert contents produce N=140 a huge number of purposes of enthusiasm with descriptors of measurement 144.And the database motor needs to seek a hopeful fingerprint among all components or expert fingerprints which are not parallel stable. To illuminate this issue, the fingerprint coordinating procedure depends on closest neighbor seek strategies. It implies that we don't seek the record reference with the same parallel representation

as the hopeful, however the file reference which gives the nearest distance amongst competitor and reference fingerprints. The primary test is to proficiently address the tradeoff between discovery speed, database size and recognition exactness. A fingerprint design connected to UGC separating performs effectively if both the fingerprint era and the fingerprint database perform productively. [7]We can't in this way separate fingerprint era from the fingerprint database. Once the reference database is populated, identification of copyrighted content can begin on UGC locales. Each transferred content on UGC destinations is fingerprinted. The choice motor begins with the worldwide descriptors and, if there should arise an occurrence of non merging, naturally changes to the neighborhood descriptors.

2.7 ROBUST DATA EMBEDDING

Fingerprinting interactive media requires the utilization of robust information embedding strategies that are equipped for withstanding assaults that piracy practicenors may utilize to expel the fingerprint.Although there are numerous systems that have been proposed for embedding the data in mixed media signal, in the continuation we will utilize the spread-range added substance embedding procedure for showing the embedding of fingerprint signs into mixed media. Spread-range embedding has demonstrated robust against various sign preparing operations, (for example, lossy pressure and sifting) and assaults.

Spread-range embedding obtains the thoughts from spread-range regulation. The fundamental procedure of spread-range embedding comprises of four stages. The initial step is to recognize and register includes that will convey watermark signals. Contingent upon the application and outline necessities, the elements can be signal specimens, change coefficients, (for example, discrete cosine changes (DCT) and discrete Fourier change (DFT) coefficients) or

different elements of the media content. Next, we produce a watermark flag and tune its quality to guarantee indistinctness

Normally we build the watermark to cover a wide range and an extensive district of the content, bringing about a watermark that looks like commotion. A third step is to add the watermark to the element signal. At long last, we supplant the first component signal with the watermarked form and change over it back to the sign space to acquire a watermarked signal. The recognition procedure for spread range watermarks starts with separating highlights from a media signal being referred to. At that point the similitude between the components and a watermark is analyzed to decide the presence or nonappearance of the watermark in the media signal. Ordinarily, a relationship likeness measure is utilized, regularly in conjunction with pre-preparing, (for example, brightening) and standardization.

Chapter 3

PROPOSED DUPLICATE IMAGE DETECTION METHOD USING DIGITAL FINGERPRINTING

3.1 AHASH

Average hashing algorithm or aHash calculates the hash using the average luminous intensity of the pixels in an image. This approach crushes the image into a grayscale 8x8 image and sets the 64 bits in the hash calculated on if the average color of image is greater or smaller than the pixel's luminous intensity value.

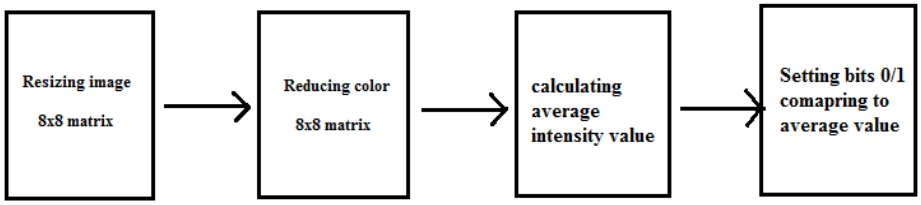

Figure 3.1 aHash steps

i. Resizing the image to a common size

At first, the image is resized to an n x n image. We used n=8 for our implementation..

ii. Grayscaling the image

By Grayscalling we reduce the color density of the image .The image is then reduced to gray scale to reduce the color information density of the image. This step changes the image's 8x8 RGB values into their single luminous intensity value with 8x8 field in the range of grayscale.

iii. Averaging the color intensity of whole matrix

The average color of the image is calculated by adding the luminous intensity values of all pixels in the image ,then dividing it by the pixel strength of the image.

iv. Hash generation

If the pixel's luminous intensity value is greater than the average color value of the image, then hash generated is 1 for the particular pixel, otherwise it is 0.Thus we get an 64 bit long binary image hash of an image through aHash.

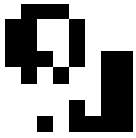

Figure 3.2: Scaled fingerprint image through aHash (8.8 pixels)

3.2 DIFFERENTIAL HASH OR DHASH

To calculate the hash of an image, dHash uses a difference field matrix.

i. Resizing the image

At first, the image is changed into an (n+1)xn image. We reduced our images into 9x8 images. Since we are going to use a difference field, we have taken the width 1 pixel greater than the height i.e. 9x8 pixels. All the high frequencies and details of the image are reduced in this step. This step ensures that hash value will not be affected by further resizing or stretching.

ii. Grayscaling the image

By Grayscalling we reduce the color density of the image .The image is then reduced to gray scale to reduce the color information density of the image. This step changes the image's 9x8 RGB values into their single luminous intensity value with 9x8 field in the range of grayscale. i.e a white pixel with RGB value (255,255,255) will be changed to it's luminous intensity value 255.

iii. Generating a difference field

The dHash algorithm works on the relative difference between the color intensity value of two neighbour pixels. . The difference field is then calculated from the 9x8 image . This identifies the relative gradient direction. In this case, after comparing each neighbour pixels the 9x8 matrix produces a 8x8 matrix. This is called as the difference field matrix of an image for calculating it's hash value.

iv. Generating the hash

For each pixels, hash value is generated based on if the left pixel's intensity value is greater than the right pixel or not. As long as we are consistent, the ordering of the comparison does not matter. Each bit in the 8x8 difference field is assigned 1 if it's the left pixel is brighter

than the right pixel, otherwise it's assigned 1. In this way an image hash of 64 bit is created from the image.

Figure 3.3: Scaled Fingerprint image through dHash(8x8 pixels)

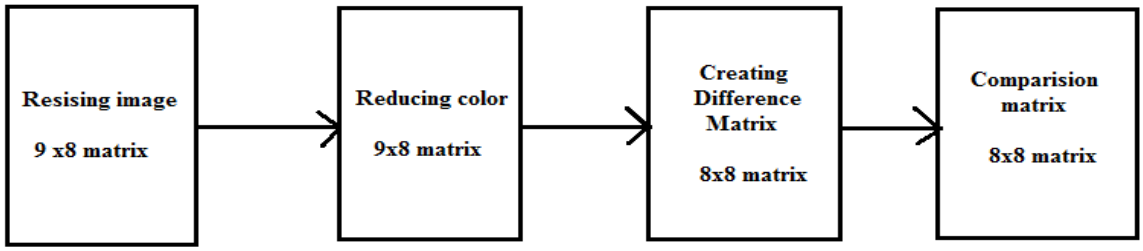

Figure 3.4 dHash steps

3.3 PERCEPTUAL HASH USING DCT (PHASH)

Average hash may be fast and easy to implement but it may fail in case of complex colluding attack. For instance if the attack includes gamma correction or even if a color histogram, aHash will generate a great number of false negatives. This is due to frequency

variations in the image. However, pHash used a more robust and complex algorithm. In our implementation, the high frequencies were reduced using a discrete cosine transform(DCT II) and then the average approach was extended to extreme.

i. Reducing the image size

Like all image hashing algorithms, pHash is also implemented using small size images. In order to simplify the computation for discrete cosine transform we have used 32x32 image which is larger than our implementations. By doing so, all the high frequencies were reduced from the image.

ii. Reducing the color of the image

To further reduce the number of computations and to get a single intensity value for each pixel, the image is converted into a grayscale image.

iii. Computing discrete cosine transform(DCT)

Image is separated into a collection of scalars and frequencies. A DCT matrix was generated and it was multiplied by the image. Then the resulting matrix was multiplied with the transpose of DCT. We have implemented this algorithm using a 32x32 DCT.

iv. Reducing the DCT further

Since the top and left 8x8 coefficients denote the lowest frequencies of an image, out of 32x32 DCT coefficients, only those were kept.

v. Computing the average DCT coefficient value.

The first DCT coefficient was radically distinct from all the other DCT coefficient. Including this DCT coefficient while averaging all the DCT values may result in an improper average DCT value. So the average was calculated just like aHash,but the first DCT coefficient was left out while calculating the average. Low DCT coefficients calculated using this algorithm are more robust to distortions since the solid color information(1^{st} DCT coefficient) was excluded during implementation.

vi. Reducing the DCT matrix further

All the hash bits were set either 1 or 0 on the basis that if the DCT value is greater than the average DCT value all 64 calculated, then it's hash bit is set as 1,or else it is set as 0. So the resulting matrix does not contain the actual low frequencies of the image, rather it is just a relative comparison matrix with the mean of all 64 DCT coefficient values. This process will be robust to all gamma and color histogram distortions provided that the structure of the image is unchanged.

vii. Final hash construction

At last the 8x8 matrix containing hash bits are set into 64-bit binary number. As long as we remain constituent the ordering does not matter. Since the comparison was done between the low frequencies, this algorithm proved more robust to distortions.

Figure 3.5: Scaled fingerprint image through pHash(8x8 pixels)

3.4 IMAGE HASH SEARCHING IN A DATABASE

The robustness and speed while matching images in a database depends on a number of factors. Complexity hash functions and size of the dataset are the most basic one of them. On the other hand the size of the dataset is purely dependent on how it was constructed. If the dataset contains only report images, then it will surely be of smaller size than the dataset containing images which were constructed while crawling and browsing through webpages. On the other hand a dataset of clicked pictures will have a greater size than a dataset containing computer graphic images. Mainly there are two ways of searching a database for images. They are listed below.

i. Logarithmic search

Since in this type of search as the dataset size increases, the time also increases logarithmically. This mi more than useful since logarithm of time makes it possible to search a huge database to find the match for a single file. However this search can only match the images which are completely identical to the original image. i.e. only images having match by 0 hamming distance will be matched. Near supplicate images can not be matched using this search.

ii. Linear search

Linear search uses the concept of hamming distance. Even a small distortion in the image will change it's hash value by a small portion. In logarithmic search we couldn't match these images. In linier search at first the hamming distance is calculated by XOR ing the input image

file hash and the registered original hash. This gives us the number of bits by which the hashes are different from each other. If the outcome of XOR is less than the predetermined threshold value, then the image is matched. However this search is very inappropriate for large datasets since the input image hash is XOR ed with all the registered hashes till it finds a match making the process bery slow.

3.5 METHOD

To test the robustness of aHash, pHAsh and dHash the following tests were carried out on image dataset.

i. Selection of dataset

A dataset containing 229 strongly attacked images named 'Copydays dataset' was selected from INRIA dataset. For our purposes this dataset contains smaller size images. All images were jpg format.

ii. Selection of Hashing algorithm

All the above mentioned image hash algorithms were run one by one on the dataset for indexing the original image files. As we know aHash is the simplest of the all and dHash is slightly more complicated than aHash. Phash uses DCT,so it is expected to be most robust of the three. Hash size and hamming distance for all the algorithms are same.

iii. Manipulating image

For testing all the below image manipulations were done using adobe photoshop softwatre.

a) **Original image**

All the above mentioned image hash algorithms were run one by one on the dataset for indexing the original image files.

b) 20% Scaling

Scaled image by 20% were used to test robustness.

c) Change in color

A slight correction in color was made. To prevent high change in picture content only 10 color correction is used.

d) Adjustment of contrast

The contrast of the image was decreased by 10%.

e) Grayscaling

Grayscalling is done on the original image.

3.6 RECORDING FALSE NEGETIVE & FALSE POSITIVE

The original registered hash and the hash after manipulating the image is compared using XOR nad the average hamming distance of all the comparisons are calculated to survey the effect of manipulation and to test the robustness of the algorithms. The matches having zero distance matches are also counted. The images which are copy of one another but are not matched as a copy by the hashing algorithm are counted. These are known as false negatives are counted. The hamming distance threshold is kept at 8 for all tests.

In order to test the robustness of the algorithm all the false positive matches are also calculated. False positives are the image matches which are not actually copied but the algorithm reports the image to be copied or colluded.

The detection rate of the proposed framework which uses image hashes mainly depends on the following factors::

- **Dataset size:** since we are mostly using linier search for comparing matches, larger the dataset size, greater will be the number of false positive and false negatives and lower will be detection speed.
- **Attack definition:** Stronger attacks and distortions have the capacity to throw off the robustness and speed of detection.
- **Size of the input file:** Smaller size images will generate the proposed hashes much fater than higher size images. That is why we have tested the dataset containing only 229 images.
- **Fingerprint size of hash size:** If the size of the hashes are larger, a low no of false positives will match but the speed of the detection will definitely decrease.

Chapter 4

RESULT AND DISCUSSION

RESULT

All the results for the mentioned image hashing algorithms are listed below. The average Hamming distance denotes the average of hamming distance of all the dataset images to all data points. The higher average hamming distance denotes more distortion of the image in case of manipulation and less robustness of the hashing algorithm. If the number of zero distance matches are very high, then the exact match number will be very high which may lead to logarithmic search for large datasets.

aHash results	No change	Color Correction	Contrast Adjustment	Grayscale filtering	20% scaling
Average distance	0	0.0402	0.643	1.433	0.129
Number of zero distances	229	194	128	93	201
False negatives	0	0	17	32	0
False positives	0	0	0	0	0

Table 4.1: aHash resullts

As shown in table 4.1, 4.2, 4.3 ,the number of zero distance matches are very low (below than half).Therefore we are not able to do a logarithmic search in log(n) time. Since dHash algorithm calculates hash by comparing the adjacent pixels of the image(content dependent), it

fares well as expected. Since the manipulations done will be fairly same to each of the pixels and average intensity values are used to generate the image hash, dHash is more robust to color correction(but not to adjustment of contrast).

dHash results	No change	Color Correction	Contrast Adjustment	Grayscale filtering	20% scaling
Average distance	0	0.0623	0.857	1.391	0.258
Number of zero distances	229	207	94	71	176
False negatives	0	0	8	22	0
False positives	0	0	1	1	0

Table 4.2: dHash results

Contradicting expectations, aHash and dHash gave poor results in grayscaling test, although both algorithms use the grayscaled image for calculating image hash. This might be due to the color analysis done by photoshop algorithms before converting image into grayscale(all colors may not be equally weighted while grayscalling in photoshop) or because in ImageHash library image is not explicitly converted into grayscale, rather only different color channels are weighted together.

pHash is proved to be the most robust to color correction, contrast adjustment and grayscaling. But it performs poorly than expected in case of scaling test. This might be due to the change in DCT value of the image done due to the bilinear interpolation used by the image processor. After surveying the results, it seems that a combination of two different image hashing

algorithms may result a better robustness like zero distance, applying both logarithmic and linier search.

pHash results	No change	Color Correction	Contrast Adjustment	Grayscale filtering	20% scaling
Average distance	0	0.072	1.749	1.969	3.632
Number of zero distances	229	221	265	107	27
False negatives	0	0	0	1	19
False positives	0	0	0	0	0

Table 4.3: pHash results

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

Digital fingerprinting is probably the best technique to be used in digital forensics. Without watermarking's embedding limitations, it enables us to identify the copied material with speed and accuracy which are very robust to attacks and distortions. Digital fingerprints used in the described image hashing algorithms provide an excellent tradeoff between accuracy, robustness and image hash matching speed. The mentioned duplicate image detection method can be used to the filter copyrighted media files for user generated content sites. Potential applications are unlimited.

Overall the implementation and manipulation test results were satisfying. The number of exact matches are very high in case of no change in color and change in color using color correction given that the threshold hamming distance for hash matching was kept at reasonable 10.False matching(different images having same hash below threshold) results were not recorded as it is quite common in case of less complicated image hashing algorithms which do not use DCT.

However we have to keep it in mind that an universal image hashing algorithms which is robust to all tyoe of distortions is not present yet. And since digital forensics is a relatively new

and vast field, most of the present image hashing algorithms have not yet been tested to most complex colluding attacks yet. But this also provides a huge opportunity for researchers to build a better image hashing algorithm.

5.2 FUTURE WORK

To make the fingerprinting algorithm more robust to rotation and other attacks, the above hashing algorithms can be combined with other image hashing algorithms to produce better and fast results. To improve the running time of algorithms, concept of embedding can be applied with fingerprint so that one does not need to extract the fingerprint each time an image is compared.

Bibliography

[1] Wu, M., Trappe, W., Wang, Z. J., & Liu, K. J. (2004). Collusion-resistant fingerprinting for multimedia. *Signal Processing Magazine,* (*IEEE*, *21*(2), pp.15-27

[2] Milano, D. (2011, April) "Content Control: Digital Watermarking and Fingerprinting."https://www.digimarc.com/resources/docs/Rozet_wp_Fingerprinting_Watermarking.pdf

[3] Gionis, A., Indyk P., Motwani R. (1999), Similarity search in high dimensions via hashing. in Proc. VLDB, pp. 518–529.

[4] Chikkerur, S., & Ratha, N. (2005, October). Impact of singular point detection on fingerprint matching performance. In *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on* IEEE. pp. 207-212

[5] Huang, J., Shi, Y., & Shi, Y. (2000). Embedding image watermarks in DC components. *Circuits and Systems for Video Technology, IEEE Transactions on*, *10*(6), pp. 974-979.

[6] Massoudi, A., Lefebvre, F., Demarty, C. H., Oisel, L., & Chupeau, B. (2006, October). A video fingerprint based on visual digest and local fingerprints. In*Image Processing, 2006 IEEE International Conference on* IEEE. pp. 2297-2300

[7] Amsaleg, L., Gros, P., & Berrani, S. A. (2004). Robust object recognition in images and the related database problems. *Multimedia Tools and applications*, *23*(3), pp. 221-235.

[8] Cox, I. J., Kilian, J., Leighton, F. T., & Shamoon, T. (1997). Secure spread spectrum watermarking for multimedia. *Image Processing, IEEE Transactions on*, *6*(12), pp. 1673-1687.

[9] Sencar, H. T., & Memon, N. (2008). Overview of state-of-the-art in digital image forensics. *Algorithms, Architectures and Information Systems Security*, *3*, pp. 325-348.

[10] Lefèbvre, F., Chupeau, B., Massoudi, A., & Diehl, E. (2009, February). Image and video fingerprinting: forensic applications. In *IS&T/SPIE Electronic Imaging* .International Society for Optics and Photonics.

[11] Wang, N., & Doube, W. (2011, May). How real is really? a perceptually motivated system for quantifying visual realism in digital images. In*Multimedia and Signal Processing (CMSP), 2011 International Conference on*(Vol. 2). IEEE. pp. 141-149

[12] Redi, J. A., Taktak, W., & Dugelay, J. L. (2011). Digital image forensics: a booklet for beginners. *Multimedia Tools and Applications*, *51*(1), pp. 133-162.

[13] Gionis, A., Indyk, P., & Motwani, R. (1999, September). Similarity search in high dimensions via hashing. In *VLDB* (Vol. 99, No. 6) pp. 518-529.

[14] Nixon, M., Nixon, M. S., & Aguado, A. S. (2012). *Feature extraction & image processing for computer vision*. Academic Press.

[15] Krawetz N.(2013).Kind of Like That. Retrieved 20 Apr. 2015 http://www.hackerfactor.com/blog/index.php?/archives/432-Looks-Like-It.html